Robust Fundamental Frequency Estimation in Coloured Noise

Esquivel Jaramillo, Alfredo; Jakobsson, Andreas; Nielsen, Jesper Kjær; Christensen, Mads Græsbøll

Published in:
2020 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2020

DOI (link to publication from Publisher):
10.1109/ICASSP40776.2020.9053018

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

ROBUST FUNDAMENTAL FREQUENCY ESTIMATION IN COLOURED NOISE

Alfredo Esquivel Jaramillo[⋆]*, Andreas Jakobsson*[†]*, Jesper Kjær Nielsen*[⋆]*, and Mads Græsbøll Christensen*[⋆]

[⋆] Audio Analysis Lab, CREATE, Aalborg University, email: {aeja, jkn, mgc}@create.aau.dk
[†] Dept. of Mathematical Statistics, Lund University, email: aj@maths.lth.se

ABSTRACT

Most parametric fundamental frequency estimators make the implicit assumption that any corrupting noise is additive, white Gaussian. Under this assumption, the maximum likelihood (ML) and the least squares estimators are the same, and statistically efficient. However, in the coloured noise case, the estimators differ, and the spectral shape of the corrupting noise should be taken into account. To allow for this, we here propose two schemes that refine the noise statistics and parameter estimates in an iterative manner, one of them based on an approximate ML solution and the other one based on removing the periodic signal obtained from a linearly constrained minimum variance (LCMV) filter. Evaluations on real speech data indicate that the iteration steps improve the estimation accuracy, therefore offering improvement over traditional non-parametric fundamental frequency methods in most of the evaluated scenarios.

Index Terms— fundamental frequency, coloured noise, maximum likelihood, pre-whitening, least-squares, LCMV filter

1. INTRODUCTION

The problem of estimating the fundamental frequency (a.k.a. pitch) of a periodic signal has received considerable attention during recent decades, and is of particular importance in many forms of audio and speech processing, such as speaker identification [1], audio coding [2], music transcription [3], and speech decomposition [4]. As opposed to correlation-based methods (e.g. YIN [5], RAPT [6]), parametric estimators [7] exploit a parametric model of the signal structure, which, if correct, allows for estimators that are more robust and that offer better resolution [8]. Many forms of parametric estimators, e.g., those based on subspace orthogonality [9], assume that the additive noise is white and Gaussian distributed (WGN), something that is rare in practice. A common consequence of this is that the found estimate is a rational number of the actual fundamental frequency when they are applied in practical noise scenarios, causing so-called octave errors. This effect may be alleviated by taking the spectral shape of the additive noise into account, which can, for example, be done by modelling the noise as an autoregressive (AR) process. Formulated mathematically, the problem may thus be expressed as follows: A set of harmonically related sinusoids, with frequencies $\{\omega_l\}$, are assumed to be observed corrupted by an additive AR noise, $e(n)$, for $n = 0, ..., N-1$, such that

$$x(n) = s(n) + e(n) = \sum_{l=-L, l\neq 0}^{L} \alpha_l e^{j\omega_l n} + e(n), \qquad (1)$$

where L is the number of harmonics and $\alpha_l = \alpha_{-l}^*$ denotes the complex amplitude of the lth harmonic. For voiced speech segments, it is often assumed that the harmonics are exact integer multiples of the fundamental ω_0, i.e., $\omega_l = \omega_0 l$, leading to the so-called harmonic model. Under the assumption that the additive noise may be well modelled as an AR process, it further holds that

$$e(n) = -\sum_{i=1}^{P} a_i e(n-i) + w(n), \qquad (2)$$

where $\{a_i\}_{i=1}^P$ are the noise AR parameters and $w(n)$ is a driving zero-mean WGN process with variance σ_w^2.

Regrettably, jointly estimating the parameters detailing both the speech $(\{\omega_l\}, \{\alpha_l\}, L)$ and the noise $(\{a_i\}, \sigma_w^2)$ is computationally prohibitive, being a multimodal and multidimensional optimization problem [10], although, reminiscent of the mixed-spectrum estimation problem presented in [11], the problem described herein may be solved in a cascaded approach, where the sinusoidal parameters and the AR noise parameters are estimated separately. However, the problems differ in two significant ways. Firstly, in [11], the signal is assumed to consist of independent sinusoids (i.e., not harmonically related) in AR noise, whereas we strive to exploit the harmonic structure of the sinusoids to allow for improved estimates [12]. Secondly, in [11], a single iteration of the procedure was sufficient for convergence, since estimating independent sinusoids under the WGN assumption is asymptotically efficient, even for coloured noise [13] but not for fundamental frequency estimation.

In the problem considered herein, estimating ω_0 without taking the AR structure into account will increase the risk of selecting an erroneous peak as the estimate, causing the noted octave error [14], and from the above discussion it is suggested that the estimates of the noise and signal parameters should rather be done in an iterative manner. This may be done by first estimating the sinusoidal frequencies without exploiting the harmonic structure, which could then be incorporated using a weighting reminiscent of the extended invariance principle (EXIP) [15]. An alternative, which is examined here, is to first form an estimate of the noise shape, and then use this in a pre-whitening step prior to estimating ω_0 (such a filtering step will not change the frequency content of the signal, merely the corresponding amplitudes [14]). However, in order to allow for reliable estimates, accurate noise AR parameters are required. For this purpose, accurate noise statistics are needed and this topic has attracted significant interest, for instance in classical algorithms such as minimum statistics (MS) [16] and minimum MSE based on speech presence probabilities (MMSE-SPP) [17], both which perform well when the noise is fairly stationary. However, for non-stationary noise types, such as babble noise, the noise parameters accuracy and the pre-whitening performance can be improved by taking into account prior spectral information on the AR-parameters of speech and noise sources [18, 19]. In this paper, extending upon the work in [11, 19, 14], we investigate two schemes for reducing the

This work was funded in part by CONACYT, under the grant 418437, and by the Swedish Research Council.

likelihood of octave errors using an iteratively refined pre-whitening filter. Both proposed methods are based on estimating the error sequence, from which a new pre-whitening filter may then be directly obtained.

2. MODEL, PROBLEM AND PROPOSED ESTIMATOR

To introduce notation and properly formulate the problem, we proceed to introduce the fundamental frequency estimator along with useful matrix and vector definitions, and discuss how the noisy signal can be pre-whitened. Consider a signal segment of N samples,

$$\mathbf{x} = \begin{bmatrix} x(0) & x(1) & \cdots & x(N-1) \end{bmatrix}^T, \tag{3}$$

with $(\cdot)^T$ denoting the transpose. Then, (1) may be written as

$$\mathbf{x} = \mathbf{s} + \mathbf{e} = \mathbf{Z}_L(\omega_0)\boldsymbol{\alpha} + \mathbf{e}, \tag{4}$$

with $\mathbf{e}$ defined similar to $\mathbf{x}$, and

$$\mathbf{Z}_L(\omega_0) = \begin{bmatrix} \mathbf{z}(\omega_0) & \mathbf{z}^*(\omega_0) & \cdots & \mathbf{z}(\omega_0 L) & \mathbf{z}^*(\omega_0 L) \end{bmatrix}, \tag{5}$$

$$\mathbf{z}(\omega) = \begin{bmatrix} 1 & e^{-j\omega} & \cdots & e^{-j\omega(N-1)} \end{bmatrix}^T, \tag{6}$$

$$\boldsymbol{\alpha} = \frac{1}{2}\begin{bmatrix} A_1 e^{j\psi_1} & \cdots & A_L e^{j\psi_L} & A_L e^{-j\psi_L} \end{bmatrix}^T, \tag{7}$$

where $A_l > 0$ denotes the (real-valued) amplitude and $\psi_l \in [0, 2\pi)$ the initial phase. For a not-voiced speech segment (including unvoiced speech and pauses), the observed signal model thus reduces to $\mathbf{x} = \mathbf{e}$. Both models may be expressed jointly as $\mathbf{x} = u\mathbf{Z}_L(\omega_0)\boldsymbol{\alpha} + \mathbf{e}$, where $u = 1$ for a voiced segment, and 0 otherwise. For white Gaussian noise, the ML estimate of $\hat{\omega}_0$ is

$$\hat{\omega}_0 = \arg\max_{\omega_0} \mathbf{x}^T \boldsymbol{\Pi}_{Z_L}(\omega_0)\mathbf{x}, \tag{8}$$

where $\boldsymbol{\Pi}_{Z_L}(\omega_0) = \mathbf{Z}_L(\omega_0)\left[\mathbf{Z}_L^H(\omega_0)\mathbf{Z}_L(\omega_0)\right]^{-1}\mathbf{Z}_L^H(\omega_0)$, which depends on the (unknown) candidate model order, L. This is the estimator that we will here refer to as the NLS (nonlinear least-squares) estimator. Fortunately, (8) may be solved efficiently in an order-recursive manner [8], after which a suitable order may be selected using a model selection criteria such as the Bayesian Information Criteria (BIC) [20, 21]. The resulting estimate would only be statistically efficient if $\mathbf{e}$ was white. As we are here concerned with coloured noise, the AR noise parameters need to be first estimated and used to pre-whiten the signal using the filter

$$A(\omega) = 1 + \sum_{i=1}^{P} a_i e^{-j\omega i}. \tag{9}$$

In order to estimate the noise parameters, the noise spectral density (PSD) $\phi_e(k)$, $k = 0, 1, ..., N-1$, can be estimated using algorithms such as MS, MMSE-SPP, the parametric NMF (Par-NMF) [19], or the model-based method introduced in [18]. Using the estimated noise PSD, the noise autocovariance sequence is then estimated as $r_e(n) = \frac{1}{N}\sum_{k=0}^{N-1}\phi_e(k)\exp\left(j\frac{2\pi}{N}nk\right), 0 \le n \le P$. Finally, a Levinson-Durbin recursion of order P is applied on $r_e(n)$ to determine the $\{a_i\}_{i=1}^P$ filter coefficients. Then, this pre-whitening filter is applied to $\mathbf{x}$, and the initial $\hat{\omega}_0$ is obtained from (8).

What has been described up to now, does not involve a reestimation step, and we now proceed to detail on how the parameters are reestimated. In a first approach, using the harmonic structure, for a given $\hat{\omega}_0$, the least squares (LS) estimate of the amplitudes may be formed as [9]

$$\hat{\boldsymbol{\alpha}} = [\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{Z}_L(\hat{\omega}_0)]^{-1}\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{x}. \tag{10}$$

Using the resulting estimate, the additive noise may be estimated by removing the harmonic model contribution from the observed signal, such that $\hat{\mathbf{e}} = \mathbf{x} - \mathbf{Z}_L(\hat{\omega}_0)\hat{\boldsymbol{\alpha}}$. From this estimate, the AR noise parameters $\{\hat{a}_i\}_{i=1}^P$ may be reestimated using the *autocorrelation method* (see, e.g., [22]) of AR modeling, which may then be used to form a new pre-whitened signal vector, from which a new estimate $\hat{\omega}_0$ can be obtained. This process can then be repeated until convergence, which is here defined as when the cost function (8) between two consecutive iterations is below a given threshold value. We refer to this method as the approximate ML approach.

The second possibility is to apply an optimal filter, capable of extracting a desired periodic signal, which satisfies the harmonic model. For this purpose, we make use of the noise covariance matrix, defined as $\mathbf{R}_\mathbf{e} = E\left[\mathbf{e}\mathbf{e}^T\right]$, where $E[\cdot]$ is the mathematical expectation operator. The applied filter will be driven by the estimated fundamental frequency $\hat{\omega}_0$ and by the estimated model order $\hat{L}$. A linear filter is applied to $\mathbf{x}$ in order to extract an arbitrary signal sample $s(n-m)$, i.e.,

$$\hat{s}(n-m) = \mathbf{h}^T\mathbf{x} = \mathbf{h}^T\mathbf{Z}_L(\hat{\omega}_0)\boldsymbol{\alpha} + \mathbf{h}^T\mathbf{e}, \tag{11}$$

where $\mathbf{h} = [h_0\ h_1\ \ldots\ h_{N-1}]^T$. It is seen that the filter affects both the speech and noise components. In order to obtain a distortionless estimate of the voiced speech sample, the constraint $\mathbf{h}^T\mathbf{Z}_L(\hat{\omega}_0) = \mathbf{b}_m^T\mathbf{Z}_L(\hat{\omega}_0)$ is imposed, which implies that the harmonics of the desired signal will not be distorted. Here, $\mathbf{b}_m^T$ corresponds to the m^{th} column of the $N \times N$ identity matrix. The problem for extracting a sample of the desired periodic signal is to minimize the residual noise variance (i.e., $E\left[(\mathbf{h}^T\mathbf{e})^2\right]$) with the above constraint, i.e.,

$$\min_{\mathbf{h}} \mathbf{h}^T\mathbf{R}_e\mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^T\mathbf{Z}_L(\hat{\omega}_0) = \mathbf{b}_m^T\mathbf{Z}_L(\hat{\omega}_0). \tag{12}$$

The filter resulting from this optimization problem is the linearly constrained minimum variance (LCMV) filter [23] and is given by

$$\mathbf{h}_{\text{LCMV}} = \mathbf{R}_e^{-1}\mathbf{Z}_L(\hat{\omega}_0)\left(\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{R}_e^{-1}\mathbf{Z}_L(\hat{\omega}_0)\right)^{-1}\mathbf{b}_m^T. \tag{13}$$

The constraints of the problem can also be modified to estimate the entire speech vector as

$$\hat{\mathbf{s}} = \mathbf{H}^T\mathbf{x} = \mathbf{H}^T\mathbf{Z}_L(\hat{\omega}_0)\boldsymbol{\alpha} + \mathbf{H}^T\mathbf{e}, \tag{14}$$

which for being distortionless must satisfy $\mathbf{H}^T\mathbf{Z}_L(\hat{\omega}_0) = \mathbf{Z}_L(\hat{\omega}_0)$. This leads to the optimization problem

$$\min_{\mathbf{H}} \text{Tr}\left\{\mathbf{H}^T\mathbf{R}_e\mathbf{H}\right\} \quad \text{s.t.} \quad \mathbf{H}^T\mathbf{Z}_L(\hat{\omega}_0) = \mathbf{Z}_L(\hat{\omega}_0). \tag{15}$$

The solution of this problem is given by

$$\mathbf{H}_{\text{LCMV}} = \mathbf{R}_e^{-1}\mathbf{Z}_L(\hat{\omega}_0)\left(\mathbf{Z}_L^H(\hat{\omega}_0)\mathbf{R}_e^{-1}\mathbf{Z}_L(\hat{\omega}_0)\right)^{-1}\mathbf{Z}_L^H(\hat{\omega}_0) \tag{16}$$

It is worth noting that one may here directly use the Gohberg-Semencul (GS) formula (see, e.g., [22]) to form the matrix inverses in closed form using the already estimated noise AR parameters:

$$\hat{\mathbf{R}}_e^{-1} = \frac{1}{\sigma_w^2}\left\{\begin{bmatrix} 1 & & & 0 \\ a_1 & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_P & \ldots & a_1 & 1 \end{bmatrix}\begin{bmatrix} 1 & a_1 & \ldots & a_P \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_1 \\ 0 & & & 1 \end{bmatrix} - \begin{bmatrix} 0 & & & 0 \\ a_P & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_1 & \ldots & a_P & 0 \end{bmatrix}\begin{bmatrix} 0 & a_P & \ldots & a_1 \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_0 \\ 0 & & & 0 \end{bmatrix}\right\} \tag{17}$$

The harmonic signal is then estimated as $\hat{\mathbf{s}} = \mathbf{H}_{\text{LCMV}}^T \mathbf{x}$, yielding the noise estimate $\hat{\mathbf{e}} = \mathbf{x} - \hat{\mathbf{s}}$, from which noise AR parameters can then be reestimated. A new pre-whitening filter is applied and a new estimate $\hat{\hat{\omega}}_0$ is reestimated. As in the ML approach, a similar reiteration between estimating the noise AR parameters and estimating the fundamental frequency to drive the LCMV filtering is possible, being repeated until convergence of the cost function (8). We refer to this as the LCMV filtering approach.

In both approaches, when the not-voiced model is favored (i.e., $\hat{L} = 0$), the estimated noise vector is $\hat{\mathbf{e}} = \mathbf{x}$, and if in the next iteration the segment is still detected as not-voiced, the process is stopped for that segment.

3. EXPERIMENTAL SETUP

We now proceed to experimentally evaluate the performance of the introduced method as compared to some well-known non-parametric methods, namely YIN, RAPT, and the Cepstrum-based method introduced in [24], here denoted Cepstrum. The speech material used for evaluation is the ten sentences in the Keele database [25], resampled to 8 kHz. This database has an annotated ground truth, which corresponds to an estimate obtained using RAPT. In the evaluation, we discard labeled segments with a negative value, i.e., we only considered voiced and not-voiced segments which have certainty of the annotated values (see [25] for further details). The ground truth values were obtained for segment lengths of 26.5 ms, with a shift of 10 ms between them. The same segment length and rate are used for all the methods. The signals are corrupted by additive babble, factory, and F-16 noise types from the NOISEX-92 database [26], at iSNRs of -5, 5, and 15 dB. The iSNR indicates the level of the clean speech signal relative to the noise component in the noisy signal, i.e. iSNR $= \frac{\sigma_s^2}{\sigma_e^2}$, where σ_s^2 is the variance of the speech signal, and σ_e^2 is the variance of the noise signal.

To assess the performance, both the fundamental frequency estimation accuracy and the voicing detection are of interest. Firstly, we use gross error rate (GER) to compute the proportion of segments where both the reference and the estimated values result in a voiced segment, and differ in more than 20%. The percentage of voiced/not-voiced detection errors is known as the voicing decision error (VDE). It is desirable to have low values of both the GER and the VDE, however some estimators may have a high VDE even if they presented a low GER as many not-voiced segments could be wrongly classified as voiced, and vice-versa. Therefore, in [27], a performance measure known as the full frame error (FFE) was proposed, which considers all kinds of possible errors: GERs, not-voiced segments wrongly classified as voiced, and voiced segments misclassified as not-voiced.

The ω_0 estimation is done on the interval [60,400] Hz for all methods. For the NLS estimator, a maximum model order of $L = 27$ is used, and the not-voiced case, i.e., $L = 0$ is considered as well. To allow that the fast NLS estimator yields accurate estimates, the signal is first pre-whitened. The AR pre-whitening order in (9) is set to $P = 25$. The applied pre-whitener is the one based on the parametric-NMF noise PSD estimate described in [19], for which a dictionary that contains typical speech and noise spectral envelopes is required. To build the dictionary, speech and noise codebooks were trained offline using a standard vector quantization technique (i.e., the Lloyd algorithm) [28]. The training is done on LSF coefficients on segments of 26.5 ms duration, with a time shift between segments of 10 ms. The quantized LSF coefficients are converted back to linear prediction coefficients of order 12. Once the speech and noise codebooks are obtained, the spectral envelopes corresponding to each codebook entry can be arranged as columns of the dictionary matrix (as described in [19]). In our case, a speech codebook of 32 entries was trained on 54 minutes of several sentences from the CMU Arctic database [29], from 4 different speakers (2 female and 2 male), resampled from 16 to 8 kHz. A noise codebook of 16 entries was trained on samples from the NOISEX-92 database of babble, F-16, factory, and street noise, resampled at 8 kHz. It is important to note that the noise samples used for the training are not the same ones used for the evaluation, and also that the speech codebook involves different speakers from the evaluation.

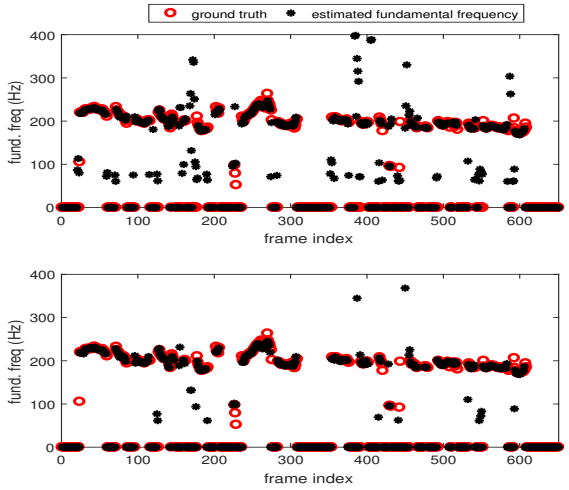

Fig. 1. Fundamental frequency ground truth and estimates without (top) and with (bottom) the proposed LCMV filtering iteration scheme.

4. EXPERIMENTAL RESULTS

We first demonstrate that the proposed reiteration scheme is able to correct wrong initial estimates. Figure 1 illustrates the ground truth and the estimated fundamental frequencies of 650 overlapping segments (approx 6.5 s) of a female speech signal excerpt of the Keele database. The clean signal is added factory noise at an iSNR of 15 dB. The ground truth is plotted in red, where a value of 0 corresponds to a not-voiced segment. In the top figure, the estimates which were obtained from the NLS estimator (after applying the pre-whitening), without the reiteration steps, are displayed. It may be seen that many segments which are not-voiced are wrongly estimated as voiced. Applying reestimation using the LCMV filtering technique (bottom figure), one may note that many of those segments are now correctly detected as not-voiced.

Next, the performance as a function of the iSNR is investigated, computed using 6 Monte-Carlo simulations, for each noise type, at each iSNR and for each one of the Keele files. The results for the three noise types in terms of GER, VDE, and FFE are shown in Figure 2, including with the corresponding 95% confidence interval. As YIN does not perform voicing detection, it is here coupled with the voicing decisions of the summation of residuals harmonics (SRH), as was also done in [30]. The NLS-NMF notation implies no reestimation, where ω_0 is estimated only one time from (8), after the pre-whitening filter from (9) is applied. The NLS-NMF Iter1 and Iter2 notation correspond to the iterative scheme based on the approximate ML approach and the LCMV filtering approach, respec-

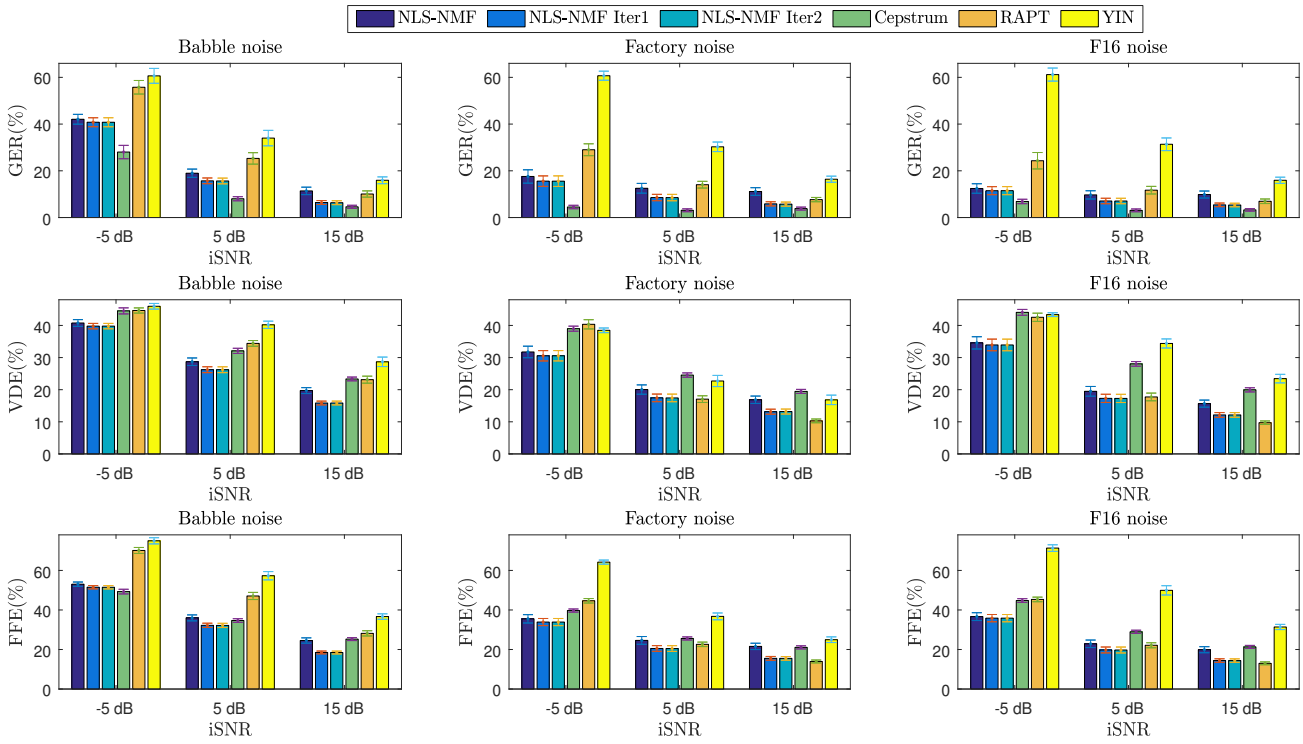

Fig. 2. The GER, voicing detection errors and total frame error of the estimated fundamental frequency, for different SNRs, for the Keele database with different noise types.

tively. We have found that convergence in both approaches typically requires 4 to 5 iterations for a voiced segment and 2 to 3 for a not-voiced one. It is important to point out that these approaches result in independent ω_0 estimates between all segments, as opposed to the other methods which include a final step of refinement, using, for instance, dynamic programming or a best local estimate selection.

First, it is noted that both presented iterative schemes result in similar performance. Furthermore, the improvements from applying the reiteration step are more evident at higher SNRs (i.e., 5 and 15 dB), as the confidence intervals are not overlapping such as in the -5 dB case. Next, it is observed that the Cepstrum method presents the lowest GER, although it results in higher voicing detection errors than the NLS estimator, even if the reiteration is not applied. Both the YIN and RAPT results are worse in terms of GER than the NLS estimator, even without the reiteration, at -5 and 5 dB. RAPT seems to better in terms of GER at 15 dB compared to NLS if the reiteration is not applied, however the performance from the approximate ML and the LCMV filtering reestimation is improved. Lower voicing detection errors are seen from the proposed methods under babble noise conditions, even without the reestimation, as compared to the three non-parametric estimators. The NLS method, even without the re-estimation, also has lower VDE as compared to YIN, in the F16 and factory noise scenarios. Comparing to RAPT, the proposed methods (with and without the re-estimation) have lower VDE at -5 dB, for both F16 and factory noise. However, at 5 and 15 dB, similar voicing detection errors are observed when using the re-estimation schemes. It is important to remember that RAPT makes use of a final dynamic programming stage, which also takes the neighbor values into account, which is not the case for the NLS estimator. In terms of full frame errors, in babble noise conditions, the proposed methods, even if there was no reiteration, have better performance than RAPT and YIN. Similar performance to the Cepstrum method is seen at -5 dB, while at 5 dB and 15 dB, the performance of NLS with reestimation is better. For factory and F16 noise scenarios, the proposed reiteration scheme yields lower FFE as compared to Cepstrum and YIN, at all SNRs, and also compared to RAPT at -5 and 5 dB. It may be noted that RAPT seems to be slightly better at 15 dB, although it should be recalled that the ground truth estimates were obtained with that method.

5. DISCUSSION

This paper considered the topic of fundamental frequency estimation in coloured noise scenarios. Most estimators make an implicit assumption that the corrupting noise is an additive, white Gaussian process, for which case the least squares estimate is statistically efficient. In practice, the additive noise shape should be taken into account in order to avoid octave errors, which may be done using a pre-whitening scheme using the estimated noise parameters. In this work, we do so by forming an AR model for the noise corrupting the speech segments, allowing us to form the required pre-whitening filter. By then estimating the harmonic components, the estimate of the additive noise may be improved, allowing for an improved pre-whitening filter, which in turn allows for an improved pitch estimate. By iteratively refining the estimates in this manner, one may reduce the risk of octave errors noticeably. Evaluated on measured speech data, we conclude that the NLS estimator reduces the number of full frame errors in most of the scenarios and therefore can offer better performance than the state-of-the-art non-parametric estimators, although only when the reiteration scheme is applied. Even without taking the correlation of consecutive estimates into account (i.e., tracking capabilities), the proposed method is more robust to the noise.

6. REFERENCES

[1] A. O. T. Hogg, C. Evers, and P. A. Naylor, "Speaker change detection using fundamental frequency with application to multi-talker segmentation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5826–5830.

[2] E. Vincent and M. D. Plumbley, "Low bit-rate object coding of musical audio using Bayesian harmonic models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1273–1282, May 2007.

[3] N. Kroher and E. Gómez, "Automatic transcription of flamenco singing from polyphonic music recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 901–913, May 2016.

[4] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "On optimal filtering for speech decomposition," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 2325–2329.

[5] A. D. Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[6] D. Talkin, "A robust algorithm for pitch tracking," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[7] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "Robust Bayesian pitch tracking based on the harmonic model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1737–1751, Nov 2019.

[8] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Processing*, vol. 135, no. Supplement C, pp. 188 – 197, 2017.

[9] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, 2009.

[10] S. M. Kay and V. Nagesha, "Maximum likelihood estimation of signals in autoregressive noise," *IEEE Transactions on Signal Processing*, vol. 42, no. 1, pp. 88–101, Jan 1994.

[11] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with applications to target feature extraction," *IEEE Transactions on Signal Processing*, vol. 44, no. 2, pp. 281–295, Feb 1996.

[12] F. Elvander, J. Ding, and A. Jakobsson, "On Harmonic Approximations of Inharmonc Signals," *arXiv:1910.07016*, 2019.

[13] P. Stoica, A. Jakobsson, and J. Li, "Cisoid parameter estimation in the colored noise case: asymptotic Cramér-rao bound, maximum likelihood, and nonlinear least-squares," *IEEE Transactions on Signal Processing*, vol. 45, no. 8, pp. 2048–2059, 1997.

[14] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "A study on how pre-whitening influences fundamental frequency estimation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6495–6499.

[15] P. Stoica and T. Söderström, "On reparametrization of loss functions used in estimation and the invariance principle," *Signal processing*, vol. 17, no. 4, pp. 383–387, 1989.

[16] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[17] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[18] J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, and J. Boldt, "Model-Based noise PSD estimation from speech in non-stationary noise," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5424–5428.

[19] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "Adaptive pre-whitening based on parametric NMF," in *2019 27th European Signal Processing Conference (EUSIPCO)*, September 2019.

[20] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, and S. H. Jensen, "Bayesian model comparison with the g-prior," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 225–238, Jan 2014.

[21] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, July 2004.

[22] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Prentice Hall, Upper Saddle River, N.J., 2005.

[23] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[24] A. M. Noll, "Cepstrum pitch determination," *The journal of the acoustical society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[25] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH*, 1995.

[26] A. Varga and H. JM Steeneken, "Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[27] W. Chu and A. Alwan, "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3969–3972.

[28] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on communications*, vol. 28, no. 1, pp. 84–95, 1980.

[29] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[30] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.